Parallel Sequential Estimation of Quantiles During Steady State Simulation

Mirko Eickhoff
Axxom Software AG, Munich, German,
www.axxom.com

Krzysztof Pawlikowski
Computer Science, University of Canterbury, Christchurch, New Zealand,
www.cosc.canterbury.ac.nz

Don McNickle
Management, University of Canterbury, Christchurch, New Zealand,
www.mang.canterbury.ac.nz

October 12, 2011

Abstract

Simulation results are often limited to mean values, even though this provides very limited information about the analyzed systems' performance. Quantile analysis provides much deeper insights into the performance of simulation system of interest. A set of quantiles can be used to approximate a cumulative distribution function, providing full information about a given performance characteristic of the simulated system. In this paper, we will present two methods for parallel sequential estimation of steady state quantiles. The quantiles are estimated using simulation output data from concurrently executed independent replications. They are calculated sequentially and on-line, i.e. during simulation, to ensure that the results are produced to a specified accuracy. The set of quantiles to be estimated can be automatically determined using efficient estimation as a criterion.

1 INTRODUCTION

Results from discrete event simulation studies are often limited to mean values, even though this provides very limited information about the analyzed systems performance. Much more meaningful insight into the performance of the system of interest is provided by quantiles, especially if several quantiles can be estimated simultaneously. For example, $q = 0.90$ or 0.95 quantiles are often specified by decision makers as the criteria of quality when considering delays or overflow probabilities in manufacturing, customer

			Quantiles					
	Mean Waiting Time		$q = .9$		$q = .99$		$q = .999$	
	10%	5%	10%	5%	10%	5%	1%	5%
$M/E_2/1$	126,430	505,721	194,624	778,497	513,317	2,053,428	2,270,981	9,083,925
$M/M/1$	170,268	681,072	259,703	1,041,677	681,130	2,732,035	3,007,636	12,063,720
$M/H_2/1$	577,022	2,308,090	836,386	3,345,344	2,122,995	8,491,982	9,276,902	37,107,609

Table 1: Number of observations for a specified relative error of estimates of the mean, and the 0.9, 0.99, and 0.999 quantiles, of the waiting time (The hyperexponential (H_2) distribution used here has a coefficient of variation $(C.V.)^2$ of 5.)

service, emergency response or telecommunication networks. The complexity and computational requirements for quantile estimation are, however, much higher than those of mean value estimation. For high traffic intensities and values of q close to 1, the approximate numbers of observations required can be estimated by using the results in [1] and [2,3]. Table 1 gives the expected number of observations that would be required for a 95% confidence interval to have relative width of 10% and 5% respectively, and for the $q = 0.9$, 0.99, and 0.999 quantiles of three queueing models, all with a traffic intensity of 0.9, in the case of $M/M/1$ queuing systems. In columns 1 and 2 these are compared to the number required to estimate the mean waiting time to the same accuracy.
Thus it can be seen that in these models estimating typical quantiles of interest can require roughly an order of magnitude more observations than the number required to estimate a mean. And further that the number of observations required can rise to extreme values when extreme quantiles are sought.
Frequently the decision maker is interested in more than one quantile at the same time, for example in estimating what the cost will be of moving from the 0.9 to the 0.95 quantile. When several quantiles are estimated from the same sample they are correlated (see Section 2.2). Thus estimation of contrasts or comparisons increases the required number of observations still further. If we wish to calculate what the effect will be of moving say from the 0.9 to the 0.95 quantile for the $M/M/1$ queue considered above, then the asymptotic correlation between these two estimates is 0.688 [???]. This in turn implies that about 5.3 million observations will be required to estimate the difference between the values of the 0.9 and the 0.95 quantile of the waiting time to a relative precision of 5% for a 95% confidence interval.
The large sample sizes required for quantile estimation strongly suggest the use of the form of parallel simulation known as Multiple Replications in Parallel (MRIP). In MRIP identical replications of the same simulation are launched either on networks of computers or multiple processor computers. Akaroa2, an automated package for launching and controlling MRIP simulations on local or global networks, is described in [4]. This can give a remarcable speedup of the execution of a simulation citeMcNickle—Pawlikowski—Ewing:2010. However when run on shared networks care must be taken to keep the communication costs low. For mean estimation, this is easily done by pooling the data. Thus only the current estimated mean and the number of observations that make it up need to be transmitted at suitably spaced checkpoints. For quantile estimation this is not possible as a sorted sample of observations must be stored and resorted as observations are added.

Here we present two sequential methods for estimation of multiple quantiles of steady state distributions in discrete-event simulation. Both methods estimate quantiles on-line, during simulation, until the stopping rule specified for sequential estimation is satisfied, for example that the estimates have reached a satisfactorily small relative error. Both methods are based on multiple independent replications and are particularly suited to the MRIP scenario. They also go some way towards keeping communications costs low by either pooling, or using only widely spaced observations. The use of independent replications ensures that at least the observations taken across replications are independent, and hence that the effect of correlation is reduced. The performance of the proposed methods is assessed analytically and empirically in a number of experiments.

Most methods of quantile estimation assume that the set of quantiles to be estimated has been selected beforehand. While our methods allow for that as well, we also consider the case of exploratory studies, when the required quantiles may not be specified in advance. We propose a method for automating the selection of multiple quantiles, using as a criterion selecting the largest possible number of quantiles with disjoint confidence intervals. Surprisingly, as a range of examples shows, this often also leads to adequate estimates of distribution functions. Thus the method also has potential for reducing the network data transmission requirements for distribution estimation. In the next section we review some basic results on estimation of quantiles and show the advantages of selecting a set of quantiles with disjoint confidence intervals. Previous works on quantile estimation in simulations is briefly reviewed in Section 2.3. This is followed by Section 3, which presents our two new methods of sequential quantile estimation using data from independent replications. In Section 4 the methods are evaluated for a range of output processes. Conclusions are given in Section 5.

2 ESTIMATION OF QUANTILES

Surveys of methods of quantile estimation in simulation output analysis can be found, in [5], [6] and [7]. In this section we briefly discuss the basics of quantile estimation. Our discussion of order statistics and

quantiles in Section 2.1 is mainly based on [8], [9] and [10].

2.1 Order Statistics and Quantiles

Let x_1, x_2, ..., x_N be a set of observations of independent and identically distributed random variables X_1, X_2, ..., X_N with common CDF $F_X(x)$. Furthermore, let $\{y_i\}_{i=1}^N$ be the ordered sequence of $\{x_i\}_{i=1}^N$, i.e. $y_1 \leq y_2 \leq \ldots \leq y_N$, and represent (ordered) sequence of random variables Y_1, Y_2, ..., Y_N. Then, Y_i is called the ith order statistic and y_i is its realisation. Because $Y_i \leq Y_{i+1}$, order statistics are dependent and not identically distributed. The CDF of Y_i is given by

$$F_{Y_i}(x) = Pr[Y_i \leq x] = \sum_{j=i}^{N} \left(\begin{array}{c} N \\ j \end{array} \right) (F_X(x))^j (1 - F_X(x))^{N-j}. \tag{1}$$

This equation allows the construction of distribution free confidence intervals for quantiles.

Let x_q, with $0 < q < 1$, be a value in the range of X, so that $F_X(x_q) = q$. Therefore,

$$x_q = F_X^{-1}(q) = \inf\{x | F_X(x) \geq q\}$$

is the population quantile of order q, if $F_X(x)$ is continuous. For simplicity, we will focus on continuous distributions only, since if $F_X(x)$ is non-continuous this definition is ambiguous.

From (1), a distribution-free confidence interval $[Y_l, Y_u]$ for an unknown value of the population quantile x_q (see [8], p.160) is

$$\begin{aligned} Pr[Y_l \leq x_q \leq Y_u] &= Pr[Y_l \leq x_q] - Pr[Y_u < x_q] \\ &\geq \sum_{j=l}^{u-1} \left(\begin{array}{c} N \\ j \end{array} \right) q^j (1-q)^{N-j}. \end{aligned} \tag{2}$$

regardless of the distribution of X.

The sample quantile $\hat{x}_q$ estimates the population quantile x_q, for a specified value of q. A common estimator is e.g.

$$\hat{x}_q = y_{\lfloor Nq+1 \rfloor}. \tag{3}$$

However, many other estimators are known. For example the weighted sum of two neighboring order statistics is another common estimator. In simulation output literature this has been discussed, for example in [11].

A sorted random sample provides natural order statistics. Therefore, we are looking for the population quantile x_q that is represented by the expected value of the ith order statistic.
$q = F_X(x_q)$ has to be estimated and x_q is given by y_i. We can see that now q depends on the form of $F_X(x)$. Thus, the bias of a general estimator of q depends on the sample size N. In [12], the properties of the approximation

$$E[Y_i] \approx F_X^{-1}\left(\frac{i}{N+1}\right) \tag{4}$$

are discussed. The error decreases with increasing sample size p and depends on derivatives of $F_X(x)$ as well as on the location of the quantile. Equation (4) suggests estimating q by $\hat{q}_i = i/(N+1)$. This estimate is asymptotically unbiased for any form of $F_X(x)$ and it is optimal for the uniform case ([8]). If the form of $F_X(x)$ is given, other superior estimators are known. As shown in [8], for the exponential case $\hat{q}_i^{(e)} = i/(N+0.5)$ and for the normal case $\hat{q}_i^{(g)} = (i-0.5)/N$ have better small sample properties than (4). However, (4) gives a general solution for the unknown distribution case and we can assume that $F_X(y_i) \approx \hat{q}_i$, if N is sufficiently large.
Following (3) or (4), one can distinguish two different approaches to quantile estimation:

Case A: observation $\rightarrow$ rank $\rightarrow$ probability: (4)

Case B: observation $\leftarrow$ rank $\leftarrow$ probability: (3)

In Case A, the order statistic y_i (resp. observation) determines the rank i, then, the probability $F_X(y_i)$ is given by a simple sample proportion $\frac{i}{N+1}$, see (4). In this case $F_X(x)$ is computed for a given x. In Case B, a probability q is given, e.g. specified by the analyst. The rank i is determined by $\lfloor Nq+1 \rfloor$ or $\lceil Nq \rceil$, see (3). The final estimate is the value of the order statistic Y_i. In this case $F_X^{-1}(q)$ is computed for a given q. Case B applies the operators $\lfloor \cdot \rfloor$ or $\lceil \cdot \rceil$, which introduce additional bias due to discontinuous sample equations. In this paper we assume that observations are collected during simulation and their probabilities need to be calculated. Thus, Case A is of higher interest and the estimator of (4) is applied, exclusively. We will compute $F_X(x)$ at a finite set of points $x_1, \cdots, x_N$.

2.2 Automated Estimaton of Multiple Quantiles with Disjoint Confidence Intervals

From our experience with Akaroa2 we have found that it is desirable to allow the option that as few as possible of the input parameters need to be specified by the user. In exploratory studies where the outcome of the simulation may be very uncertain, even selecting the particular quantiles to be estimated may be difficult. For example if a 0.95 quantile is pre-specified it may turn out to have an impractically large or small value, meaning usually that the simulation has to be run again. To avoid this situation we describe a method which allows a decision maker to specify, without prior knowledge of the distribution, that a wide range of quantiles, caited to many decision making problems, are to be estimated efficiently. This method also turns out to give surprisingly good results when used to estimate the distribution, at lower communication cost than more direct methods. The idea of using a set of qualites to estimate a distribution is known in the literature (see Section 10.4 in [4]). In Section 4 we will use sets of qualites selected in this way, to assess the quality of the proposed estimation methods.

If two or more quantiles are estimated from the same random sample their estimates will be correlated. This correlation depends on the underlying probability distribution. However for large samples $Var(\hat{x}_q) = q(1-q))/Nf(x_q)^2$, where f is the probability density function evaluated at x_q. and that $Cov(\hat{x}_{q_1}, \hat{x}_{q_2}) = q_1(1-q_2)/Nf(x_{q_1})f(x_{q_2})$, where $0 < q_1 < q_2 < 1$ [13]. Thus the asymptotic correlation for large samples and $0 < q_1 < q_2 < 1$. is:

$$\sqrt{q_1(1-q_2)/q_2(1-q_1)} \tag{5}$$

Since correlations are independent of the ordering of the observations, this limit also applies to the serially correlated observations produced from simulation, since the serial correlation can be removed by randomly reordering the observations.

Thus for a fixed value of q_1 the asymptotic correlation tends to zero for q_2 close to 1. On the other hand, for q_2 close to q_1 the asymptotic correlation tends to 1. This shows that we need a mechanism to control the correlation of two neighboring quantile estimates. Preferably they should not be located too close to each other, because then they are likely to be strongly correlated. One possible method, motivated by these observations, is discussed next.

Equation (2) allows us to construct a confidence interval for the unknown value of a population quantile x_q based on two order statistics Y_l and Y_u. $Pr[Y_l \leq x_q \leq Y_u]$ can be calculated for arbitrary ranks $1 \leq l \leq u \leq N$. It is not necessary, but could be desirable, that the confidence interval $[Y_l, Y_u]$ is symmetric with each of its halves contains half of the probability mass contained in whole interval.

Definition 1 *Let y_c be an approximately unbiased estimate of the unknown value of the population quantile x_q. The confidence interval $Pr[Y_l \leq x_q \leq Y_u] \geq 1-\alpha$ is* balanced *if*

$$\begin{aligned} Pr[Y_l \leq x_q \leq y_c] &\geq \frac{1-\alpha}{2} \quad \textit{and} \\ Pr[y_c \leq x_q \leq Y_u] &\geq \frac{1-\alpha}{2} \end{aligned} \tag{6}$$

This definition comes from the concept of mid-p confidence intervals, [14]. Other common approaches are to construct a confidence interval that has minimum width or that has $x_q - Y_l = Y_u - x_q$. However, we construct the confidence interval on basis of (6) because, in the balanced case, u and l can be calculated independently of each other.

Any of two confidence sub-intervals limited by y_c can be calculated by estimating q by $\hat{q}_i$ in the general case, or by $\hat{q}_i^{(e)}$ or $\hat{q}_i^{(g)}$, where $i = c$, in the exponential and Gaussian cases. Once q is determined

we can initialize $l = u = c$ and calculate $Pr[Y_l \leq x_q \leq y_c]$ and $Pr[y_c \leq x_q \leq Y_u]$ by applying (2), separately. l is decreased and u is increased until the conditions of (6) are met.

For balanced confidence intervals $Pr[Y_{l_1} \leq x_{q_1} \leq Y_{u_1}] \geq 1 - \alpha$ and $Pr[Y_{l_2} \leq x_{q_2} \leq Y_{u_2}] \geq 1 - \alpha$, the estimators of x_{q_1} and x_{q_2} are dependent because they are taken from the same sample. However, by choosing disjoint confidence intervals, i.e. $u_1 \leq l_2$, we can ensure that at most $\frac{\alpha}{2}$ of the probability mass of both distributions overlap. If α is made sufficiently small, e.g. $\alpha \leq 0.1$, the degree of dependence should be reduced.

Let us assume that the size N of the random sample is odd and we start by selecting y_{c_1}, with $c_1 = \frac{N+1}{2}$. The assumption of an odd N ensures that y_{c_1} is an unbiased estimate of the median $x_{0.5}$ and that all our further results are symmetric with center $x_{0.5}$. The balanced confidence interval $[Y_{l_1}, Y_{u_1}]$ at confidence level $1 - \alpha$, divided in two at y_{c_1}, can now be calculated as described above. We start with $l_1 = u_1 = c_1$ and decrease l_1 and increase u_1 until (6) holds. If this is successful, the first confidence interval $[Y_{l_1}, Y_{u_1}]$ is given. To find the second confidence interval $[Y_{l_2}, Y_{u_2}]$ with $u_1 \leq l_2$, we have to find y_{c_2} that estimates x_{q_2} so that $Pr[Y_{l_2} \leq x_{q_2} \leq y_{c_2}] \geq \frac{1-\alpha}{2}$. We start this search with $c_2 = u_1$. Now, q_2 can be determined by $\hat{q}_i$ in the general case, or by $\hat{q}_i^{(e)}$ or $\hat{q}_i^{(g)}$ for the exponential or normal case. These equations describe how to find the unknown position of the quantile that is estimated by the given order statistic. After q_2 is estimated, y_{l_2} can be calculated. If $u_1 > l_2$, this choice of c_2 should be rejected, c_2 set equal to $u_1 + 1$, and tested. The search can be stopped if $c_2 \leq N$ is violated, as no more disjoint confidence intervals can be fitted in the unprocessed area. If $u_1 \leq l_2$ holds, a valid choice of c_2 has been found and additionally u_2 must be tested. If no $u_2 \leq N$ can be found, the search can be stopped; otherwise another disjoint confidence interval has been found.

Here, we have described the search for disjoint and balanced confidence intervals for x_q with $q \geq 0.5$. The search for x_q with $q \leq 0.5$ can be done analogously. If the sample size N is even, two different starting points for $q \leq 0.5$ and $q \geq 0.5$ can be used. A flowchart for these algorithms can be found in [5].

In Table 2 the results of the algorithm are shown for $N = 999$ and $\alpha = 0.05$. The first column shows q, selected by (4). The second column is the rank c of the associated order statistic. The third and the fourth column are the bounds u and l of the balanced confidence interval. We can see that all confidence intervals are disjoint and that $\{y_i\}_{i=1}^N$ is split into 23 parts. Consecutive confidence intervals are contiguous, so that $u_{i-1} = l_i$ holds for any i. The position and size of the confidence intervals are symmetric with center at $q = 0.5$. Only the lowest and highest order statistics are not used in any confidence interval. The probability density functions of (1) and the selected order statistics are depicted in Figure 1. The density functions for low and high quantiles are asymmetric and their shapes indicate that they produce narrower confidence intervals.The confidence interval of the median is the largest and

q	c	l	u
0.007	7	2	14
0.023	23	14	34
0.047	47	34	61
0.077	77	61	95
0.114	114	95	135
0.157	157	135	181
0.206	206	181	232
0.259	259	232	287
0.316	316	287	346
0.376	376	346	407
0.438	438	407	469
0.5	500	469	531
0.562	562	531	593
0.624	624	593	654
0.684	684	654	713
0.741	741	713	768
0.794	794	768	819
0.843	843	819	865
0.886	886	865	905
0.923	923	905	939
0.953	953	939	966
0.977	977	966	986
0.993	993	986	998

Table 2: Disjoint and balanced confidence intervals for $N = 999$ and $\alpha = 0.05$.

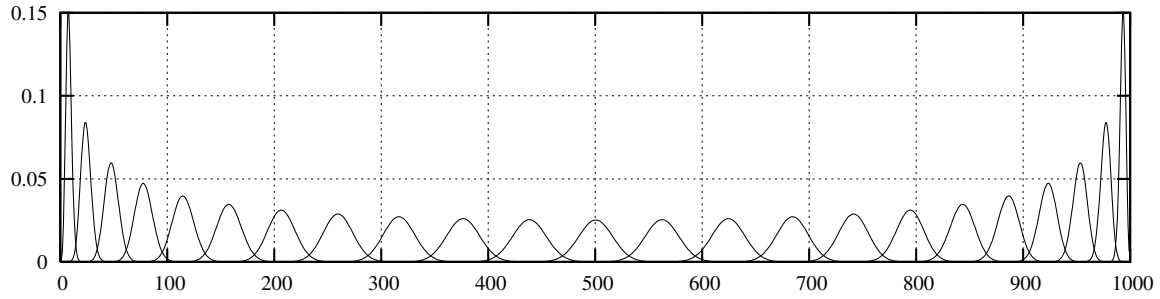


Figure 1: Density functions of quantiles with balanced and disjoint CIs, for $N = 999$ and $\alpha = 0.05$.

its density function is symmetric. Thus, with its concentration on the tails of the distribution, the set of quantiles should suit many decision-making problems.

Because $l = 469$ and $u = 531$ the confidence interval of the median contains $u-l = 62$ order statistics. Using (4) we obtain a confidence interval size of $\frac{u-l}{N+1} = \frac{531-469}{999+1} = 0.062$ in the probability domain. The maximum size can be controlled by a specified threshold in the probability domain. It is possible to calculated how large the sample size N has to be to satisfy this threshold without the knowledge of Y_l and Y_u. A larger N leads to a smaller distance $u - l$. The confidence interval size in the domain of the measure can be calculated by $Y_u - Y_l$, which depends on the underlying distribution $F_X(x)$. In this domain, the confidence interval of the median is not necessarily the largest. Again, the maximum error can be controlled by a specified threshold in the domain of the measure. In this domain, controlling the error is more difficult because the values of the order statistic Y_l and Y_u are needed. If $Y_u - Y_l$ is larger than the threshold a sample of larger size has to be collected. This should be done sequentially until $Y_u - Y_l$ is small enough to meet the threshold.

2.3 Estimation of Quantiles from Simulation Output

Here we give a brief review of previous work on quantile estimation in simulation. In previous sections we assumed random samples that are independent and identically distributed. However, in general the output stream of a simulation is a stochastic process whose states at different time instances are dependent and not identically distributed. Thus, in general, the assumption of independent and identically distributed data does not hold. Most authors have assumed instead the Φ missing condition [15] which limits the dependence between widely spaced observation and quantile estimator is at least asymptotically unbiased and has its variance decreases to zero as the number of observations increases and we shall do the same.

In quantile estimation storage requirements and calculation time are also important factors because as Table 1 shows a huge amount of output data may be needed to obtain trustworthy results. Therefore, not only the mathematical definition of the estimator, but also the way it is computed is of interest since only efficient data structures and algorithms can be practically applied.

The estimation of one *single quantile* is usually done to analyze the tail behavior of a distribution. In this case typically the 0.95-quantile (resp. 0.05-quantile) is estimated. The estimation of a single quantile of the steady state distribution, when simulating a single instance of a time-stationary process, has been considered for example by [16], [17], [18], [19], [20] and [6]. The methods of Iglehart [16] and Seila [17] are limited to regenerative processes, using the subdivision of the output data into its regenerative cycles as a natural way to overcome the problem of autocorrelation. The method of Seila extends that of Iglehart by grouping the regenerative cycles into batches. The number of parameters which have to be specified by the user is reduced by this batching approach to one: the batch size. However, the determination of the batch size is a problem in itself, common to every batching approach in simulation. To choose an appropriate value is difficult for an inexperienced user. The method of Heidelberger and Lewis [18] is not limited to regenerative processes. Their point estimator based on ordered data can be used with autocorrelated samples, although autocorrelations would inflate its variance, leading to a larger interval estimate. It can be calculated using a spectral method (see [21]). Alternatively, the original sample of correlated data can be transformed in a secondary sample of almost independent data

by using a batch means method (see e.g. [22]). The method of [19] uses a completely different approach. Their estimator is calculated by applying markers, which are adjusted when collecting new observations. This is done by a piecewise-parabolic interpolation. Because of this interpolation, the method is not recommended for estimation of quantiles in discontinuous distribution functions. The estimator is more complicated than the usual estimators based on ordered data. However, its principal advantage is that the calculations require only a constant (and small) amount of memory. [20] describe a method that estimates a quantile by focusing on observations which are located in the neighborhood of this quantile. Their method is sequential to ensure accurate final estimates. A method for quantile estimation in finite-horizon simulation is described in [23] and [24]. This method is based on multiple replications of the finite-horizon simulation. Additionally, these replications are made to be negatively correlated with each other to reduced variance of the estimates (see also [25]).

If the analyst is also interested in the complete distribution function of a performance measure the estimation of *several quantiles* is useful, because the quantiles describe the probability distribution at specific points. The estimation of several quantiles of the steady state distribution is addressed in [26]. The method of [19] is extended by introducing additional markers to estimate more quantiles. The adjustment of the markers is done in the same way as before. An investigation of the variance of this method is given in [27]. A different approach is proposed in [28]. In most studies the location in the range of the measure is estimated for a fixed probability. Here, the probability of a predefined category of the range of the measure is calculated. The most obvious category is $X \leq x$, resulting in a point and interval estimate of $F_X(x)$. In addition a method of controlling the simultaneous precision of several such measures, based on the Bonferroni inequality is demonstrated.

One of the main difficulties in quantile estimation is the high computational effort and the large amount of storage needed to order the observations. Therefore, Heidelberger and Welch reduce the sample size by a maximum transformation (see [18]). Jain, Chlamtac and Raatikainen go further and avoid sorting the output data by using an interpolation. In [29], [30], [31] and [32], quantile estimators based on order statistics have been considered again, as recent technological advances (decreasing costs of memory and increasing processor speeds) have made such approach more practical. Wood and Schmeiser describe a batching method for quantiles which is similar to batch means and consider different quantile estimators, all based on ordered observations. The batch statistic is given by one of four quantile estimators, which are all based on ordered observations. Again, the difficulty is how to chose an appropriate batch size. In [31] the previous method of [20], for estimating a single quantile, is extended to the problem of several quantiles. In [33] performance of single and multiple quantile estimators is assessed. The coverage of the single quantile estimator proves to be higher than expected. This estimator also reduces the amount of data that needs to be stored. However, the average run length of all experiments is below $12.5 \cdot 10^6$ observations. Storing this data takes about 200 MB of memory, assuming 16 bit numbers, and is no problem for modern computers. Even sorting of such a data set should not take long if efficient data structures are used and the data is sorted by merging small samples into the already sorted large sample. For some quantiles of correlated data the coverage of the multiple quantile estimator is not as good as expected. This might indicate that the runs-up test (see [34]), which is used to transform correlated data into quasi-independent data, is not optimal.

Two different density estimators are described in [35]. One is based on histogram estimation, which is closely related to quantile estimation. The other one is based on the use of a kernel function. Chen and Kelton show experimentally that the histogram density estimator is superior since it produces a better coverage of estimated confidence intervals. They conclude that the histogram procedure is more suitable as a generic density estimation procedure since it requires less computation and delivers valid confidence intervals.

3 TWO METHODS FOR PARALLEL SEQUENTIAL ESTIMATION OF QUANTILES

We assume that the simulation has reached steady state. Steady state in terms of the probability distribution is given if beyond some observation index T

$$\forall(i \geq T, \Delta \geq 0, x) : F_{X_i}(x) \simeq F_{X_{i+\Delta}}(x). \tag{7}$$

By "$\simeq$" we denote closeness of distributions, for example in the Kolmogorov sense:

$$\sup_{-\infty<x<\infty} |F_{X_i}(x) - F_{X_{i+\Delta}}(x)|. \tag{8}$$

Almost all methods proposed for estimating the duration of the initial transient in simulation output data analysis, have been proposed and tested mean value analysis only. This is not adequate for our purpose, so in [36] we have proposed a new method for the estimation of T, the length of the initial transient phase based on convergence of probability distributions to steady state. This method is based on concurrent independent replications of simulation. In such a scenario, replications of the same simulation are executed in parallel, using either different pseudo random number generators, or non-overlapping sequences of pseudo-random numbers from the same generator. Thus, it fits in well with the methodology proposed in this paper and is used here.

For quantile estimation, an obvious possibility is to exploit the independence of observations taken from p independent replications, and that will be the basis of our first method, discussed in Section 3.1. The key question will be if it is currently practical to make p large enough to produce good quality estimaties.

Using p independent replications, we obtain p independent streams of observations $\{x_{1,i}\}_{i=T}^{n}$, $\{x_{2,i}\}_{i=T}^{n}$, ..., $\{x_{p,i}\}_{i=T}^{n}$ representing p sequences of random variables $\{X_{1,i}\}_{i=T}^{n}$, $\{X_{2,i}\}_{i=T}^{n}$, ..., $\{X_{p,i}\}_{i=T}^{n}$. (For simplicity, we assume the sequences are the same length.) The independence of all replications implies that, for fixed i, the random variables $\{X_{j,i}\}_{j=1}^{p}$ are independent of each other. Therefore, for a fixed observation index, statistical methods valid for random samples of independent and identically distributed observations are directly applicable to $\{X_{j,i}\}_{j=1}^{p}$. The definition of the population quantile has to be extended to include the observation index i:

$$x_{q,i} = F_{X_i}^{-1}(q) = \inf\{x | F_{X_i}(x) \geq q\}.$$

Let $\{Y_{j,i}\}_{j=1}^{p}$ be the ordered version of $\{X_{j,i}\}_{j=1}^{p}$ and let $\{y_{j,i}\}_{i=T}^{n}$ be its realization related with the replication j. Then, $Y_{j,i}$ represents the jth order statistic at observation index i, and n can be regarded as the time horizon of simulation, as it means the total number of observations number recorded in each simulated replication (including observations in the initial transient stage of simulation). If simulation reaches the observation index n, output data produced by p replications for the purpose of steady-state quantile estimation could be stored in a matrix with p rows and $n - T + 1$ columns.

3.1 Means of Order Statistics

Our aim is to estimate one or more quantiles of $F_{X_\infty}(x)$. A natural approach is to estimate those quantiles x_{q_j} which are represented by the order statistics $\{Y_{j,i}\}_{i=T}^{n}$. For an unknown distribution, q_j can be estimated by $\hat{q}_j$, by applying (4). The differences in distributions of X_i for $i \geq T$ are negligible, so, the mean

$$\hat{x}_{\hat{q}_j} = \frac{1}{n-T+1} \sum_{i=T}^{n} y_{j,i}, \tag{9}$$

is a point estimate of $F_{X_\infty}^{-1}(q_j)$.

Theorem 3.1 *The mean of the jth order statistic $\hat{x}_{\hat{q}_j}$ is an asymptotically unbiased estimate of $F_{X_\infty}^{-1}(q_j)$ for large p and $i \geq T$.*

Proof The expected value of (9) is

$$E[\hat{x}_{\hat{q}_j}] = \frac{1}{n-T+1} \sum_{i=T}^{n} \mathrm{E}[Y_{j,i}], \tag{10}$$

with $E[Y_{j,i}] = F^{-1}_{X_i}(q_j)$ for large values of p, see (4) and [12] or [8]. Since, all X_T, X_{T+1}, ... are assumed to be identically distributed, Equation (10) reduces to

$$\begin{aligned} E[\hat{x}_{\hat{q}_j}] &= \frac{1}{n-T+1}\sum_{i=T}^{n} F^{-1}_{X_i}(q_j) \\ &= \frac{1}{n-T+1}\sum_{i=T}^{n} F^{-1}_{X_\infty}(q_j) \\ &= F^{-1}_{X_\infty}(q_j). \end{aligned} \tag{11}$$

The estimate $\hat{x}_{\hat{q}_j}$ is asymptotically unbiased, i.e. $\mathrm{E}[\hat{x}_{\hat{q}_j}] - F^{-1}_X(q_j) = 0$, because (11) holds for large p and $i \geq T$.

To establish a confidence interval for (9) its variance $\mathrm{Var}[\hat{x}_{\hat{q}_j}]$ is needed. Note, that all $Y_{j,T}$, $Y_{j,T+1}$, ... are correlated and hence the variance cannot be estimated directly. The form of (9) is identical to mean value estimators of single simulation runs. Its special feature is that each component represents a quantile. The Φ mixing property required for the convergence of $Var[\hat{x}_{\hat{q}_j}]$ to zero clearly extends to the case of the combined streams of observations. Therefore, known techniques for estimation of variance of mean value estimators can be applied, for example spectral analysis (see e.g. [21]) and batching methods (see e.g. [37]).

Both spectral analysis and batching methods have already been suggested in [18] for variance estimation in quantile analysis. The maximum transformation is used to obtain extreme quantiles of the output process. Here, we replace the maximum transformation with (9) and extend the method to multiple independent replications. Further details on how batching and spectral analysis methods are applied can be found in [38].

Equation (9) is suitable for a sequential approach because n can be extended. Extensions of batching and spectral analysis to a sequential approach are discussed in e.g. [39]. Let Δ_{q_j} be the halfwidth of the confidence interval of point estimate $\hat{x}_{\hat{q}_j}$, calculated on basis of $\mathrm{Var}[\hat{x}_{\hat{q}_j}]$ and the Student t-distribution. The stopping criterion

$$\Delta_{q_j}/D \leq \epsilon_{max}$$

can be used to stop the process of sequential estimation. D is a value which is used to standardize the halfwidth of the confidence interval and ϵ_{max} is the maximum acceptable relative error, where $0 < \epsilon_{max} \leq 0.1$. In mean value analysis D is usually the point estimate itself. However, here we have to take into account that it is quite likely that $F^{-1}_{X_i}(q_j) \approx 0$ holds for one of our quantile estimates. Furthermore, it is desirable to standardize all quantile estimates by the same value D. Therefore, we choose $D = \hat{x}_{\hat{q}_p} - \hat{x}_{\hat{q}_1}$, which is the estimated range of $F_{X_\infty}(x)$. This guarantees small limit values of relative errors for all the estimated quantiles.

3.2 Pooling Spaced Observations

Approximate independence can also be achieved by establishing a pool of observations, spaced far apart from each other. Let s be an adequate space size. Then X_T, X_{T+s}, X_{T+2s}, ... can be regarded as nearly independent. When using p replications the pool of observations is given by

$$\mathrm{P} = \{\{X_{j,T+si}\}_{j=1}^{p}\}_{i=0}^{n}$$

observations which have been filtered out from total of $p \times (n - T + 1) \times s$ observations recorded in steady-state phase of a given simulation. The size of this pool can be as large as desired and contains approximately independent and identically distributed data if T and s are large. Because of this, standard quantile estimators are directly applicable to estimate $F^{-1}_{X_\infty}(q)$.

The determination of an adequate value of s is similar to the determination of a batch size for batch means, as pooling is just a special kind of batching. For this task correlation tests are needed such as the run tests (see e.g. [40]) or permutation tests (see e.g. [41]). The von Neumann ratio test is probably the most recommended correlation test today (see e.g. [22]). However, here we wish to estimate a set of quantiles and, thus, have to find an overall space size s that is valid for all sequences $\{X_{j,T+i}\}_{i=0}^{\infty}$, where

$1 \leq j \leq p$. Thus, we decided on a correlation test based on permutations and the median confidence interval (see e.g. [42]), which is described next.

Let $\hat{r}^{(p)}(P_1)$ be Pearson's correlation coefficient of the original lag-1 paired spaced sequence $\{(X_{j,T+is}; X_{j,T+(i+1)s})\}_{i=0}^{n-1}$, and $\hat{r}^{(p)}(P_k)$ be Pearson's correlation coefficient for the lag-1 paired data of the kth permutation of $\{X_{j,T+is}\}_{i=0}^{n}$ with $2 \leq k \leq (n!)$. In [43] the first four moments of Pearson's correlation coefficient are derived. Here, the first and the third moment are of special interest: $\mathrm{E}[\hat{r}^{(p)}] = 0$ holds even for small samples and $\mathrm{Skew}[\hat{r}^{(p)}] = 0$ holds approximately. Therefore, $F_{\hat{r}^{(p)}}(0) = 0.5$ is approximately true. The null hypothesis of our test is that $\{X_{j,T+is}\}_{i=0}^{n}$ is an independent sequence.

$$Pr[|\hat{r}^{(p)}(P_k)| < |\hat{r}^{(p)}(P_1)|] = \frac{1}{2}$$

holds under the null hypothesis for a randomly chosen permutation P_k. For K randomly chosen permutations $P_{k_1}, \ldots, P_{k_K}$ we can derive

$$Pr[\forall l(1 \leq l \leq K) : |\hat{r}^{(p)}(P_{k_l})| < |\hat{r}^{(p)}(P_1)|] = \frac{1}{2^K}.$$

On the basis of this equation the following confidence interval can be established:

$$Pr[-\Delta \leq \hat{r}^{(p)}(P_1) \leq \Delta] = 1 - \frac{1}{2^K}$$

with halfwidth

$$\Delta = \max_{1 \leq l \leq K} \left(|\hat{r}^{(p)}(P_{k_l})| \right).$$

If $\hat{r}^{(p)}(P_1)$ is not within the confidence interval, the null hypothesis must be rejected at significance level $1 - \frac{1}{2^K}$. The advantage of using this confidence interval is that the assumption of zero skewness is milder than the assumption of a Gaussian distribution. For only $K = 6$ permutations the confidence level is already > 0.95 so K can be set arbitrarily. For our purpose of estimating the overall space size s for p independent replications this correlation test is performed on $\{X_{j,T+is}\}_{i=0}^{n}$ for any j.

By adding an additional sequence $\{X_{j,T+s(n+1)}\}_{j=1}^{p}$ of previously unprocessed observations at index $n+1$ the sample size can be extended by p observations. For quantile estimation based on order statistics the sample has to be sorted. The most efficient way of sorting in this case is to merge two already sorted samples. Let assume that P of size pn is already sorted. The new sample $\{X_{j,T+s(n+1)}\}_{j=1}^{p}$ can be sorted in $O(p \log(p))$. Merging of P and the new observations can be done in $O(pn + p)$. So the total runtime of adding new observations to a sorted pool of data is $O(p \log(p) + p(n + 1))$. Because usually $n >> p$ holds, we can simplify the runtime to $O(pn)$, which is efficient.

For the multiple quantile technique described in Section 2.2, the number of quantiles selected depends on the sample size, $N = pn$. So, a stopping criterion could be defined by simply setting a minimum number of quantiles which are to be selected.

On the other hand, a stopping criterion can depend on the size of the confidence interval. Let the confidence interval $Pr(Y_l \leq x_q \leq Y_u) \geq 1 - \alpha$ be a confidence interval of the unknown value of the population quantile x_q, where Y_i is the ith order statistic from the pool of observations in P. Similar to the stopping criterion of the previous section, we can set

$$\frac{y_u - y_l}{2(y_{pn} - y_1)} \leq \epsilon_{max}.$$

Here the halfwidth of confidence interval, $0.5(y_u - y_l)$ is standardized by the range $y_{pn} - y_1$ to avoid a division by a value close to zero.

3.3 Discussion

In the quantile estimation method of Section 3.1 the size of the data sample is $p \times (n - T + 1)$. However, the size p of the secondary sample, which is used for quantile estimation, is given by the number of replications. The maximum number of parallel replications will usually be restricted by the aveilable hardware, like the number of processors in a grid computing system. If p is not large enough to fulfill the conditions of Theorem 3.1, we can expect biased quantile estimates. However, the method of Section 3.1

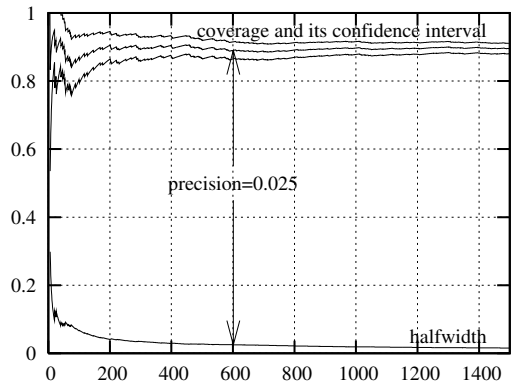


Figure 2: Quantile coverage versus the number of independent simulation experiments.

does not give a mechanism which would control correlation between quantile estimates in a set of data collected during one step. The described stopping criteria guarantees small statistical error only if Theorem 3.1 holds.

In the quantile estimation method of Section 3.2 the pooled sample P is increased until all quantiles have sufficiently small statistical error. This is an advantage compared to the previous method where the data size used for estimation is fixed at p. Another advantage of Pooling Spaced Data is that correlation between the quantile estimates can be controlled, for example by choosing disjoint confidence intervals of quantile estimates. Again, the described stopping criteria guarantee small statistical error.

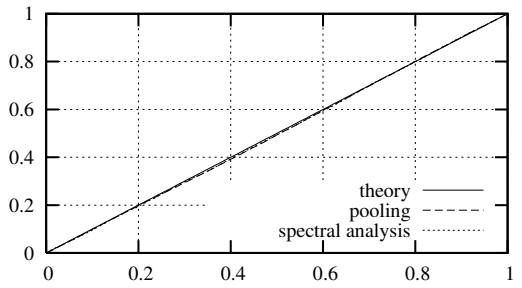
4 EXAMPLES

In this section we will test the quantile estimation methods of Section 3.1 and 3.2 on a range of examples. Rather than selecting a few quantiles for testing we use the automated method for the selection of the quantiles to be simulated, as in Section 2.2. For clarity in the graphs, the individual quantiles are suppressed in the plots and instead they are joined up with a piecewise linear curve to produce an approximation to the known CDF. As we note that in all the examples we tried this ad hoc method gave a surprisingly close approximation to the theoretical CDF, as can be seen in Figures 4 and 6. While this probably cannot compete with methods of directly estimating the CDF, we conjecture that our methods, which first produce estimates of a set of quantiles to a pre-specified relative precision, that subsequently can be used for an ad hoc estimate of the CDF may, in many circumstances, be more attractive to the decision maker than attempting to estimate the CDF across its entire range.

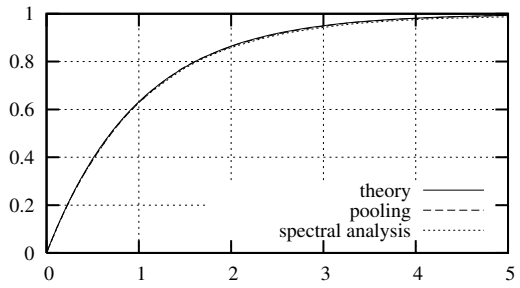
We should emphasize that the plots of quantiles in Figures 4 and 6 are the results of single simulation experiments, with random starting seeds. The plotted results were not selected in any way, and hence are representative of what could be expected in practice. The quality of the estimated quantiles assessed by means of coverage analysis of their estimated confidence intervals. That is, does an estimated 95% confidence interval for the quantile actually contain the true value 95% of the time? Coverage analysis has the advantage of capturing both any bias in the estimates, and whether or not the sequential procedures are stopping at the correct point. This is done following the steps described in [44]: the entire experiment is replicated, and in each experiment the coverage analysis is done sequentially until a certain precision is reached. Replication continues until a minimum number of "bad" confidence intervals have been detected. A bad confidence interval is one that does not contain the theoretical result. The precision is here measured by the halfwidth of the coverage's confidence interval,

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{c(1-c)}{n_c}},$$

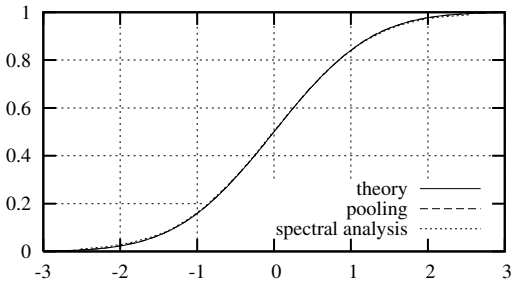
where z is the q-quantile of the standard normal distribution, c is the coverage and n_c is the number of replications conducted in a given analysis of coverage. Here, the threshold of the precision is taken to be 0.025 at the 0.95 confidence level. This ensures that the coverage convergence curves reach a stable level as Figure 2 shows.



(a) uniform: $F_{X_i}(x) = U(x; 0, 1)$



(b) exponential: $F_{X_i}(x) = Exp(x; 1)$



(c) Normal: $F_{X_i}(x) = N(x; 0, 1)$

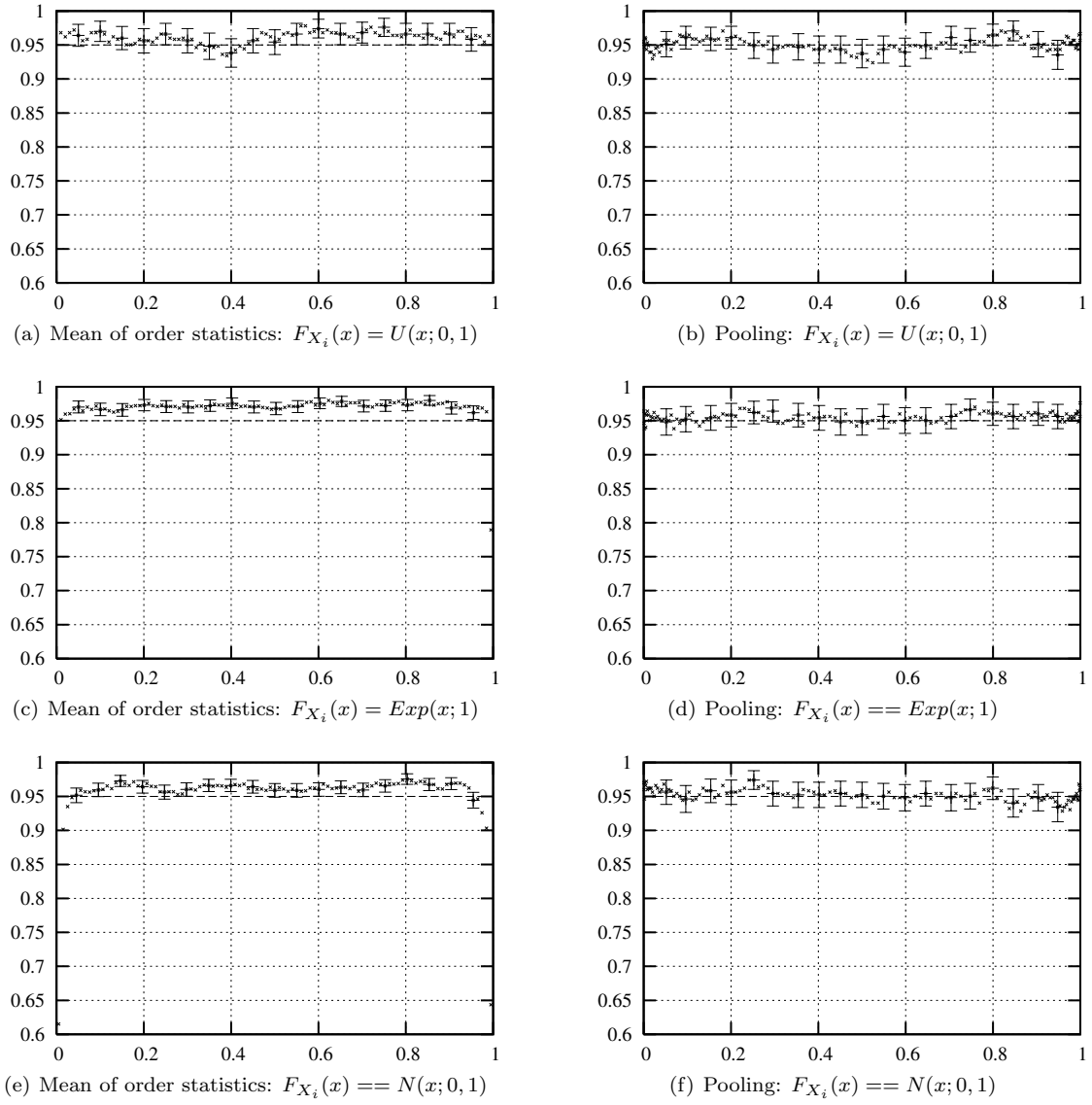
Figure 3: Expected and estimated CDFs.

4.1 Basic Processes

We start our experimental investigation with three very basic output processes. These processes do not have a transient phase and they are not autocorrelated, therefore, they provide observations which are independent and identically distributed. This allows us to test the quantile estimators themselves, independent of other parts of the proposed methods. For the method of Section 3.1 we applied the estimators $\hat{q}_i$, $\hat{q}_i^{(e)}$ or $\hat{q}_i^{(g)}$ depending on the given case. For the method of Section 3.2 we always used the estimator $\hat{q}_i$, regardless of the form of the underlying distribution. Here, estimation is done by using pooled data. In Figure 3, it can be seen that there is very little difference between the estimated and the known CDF, implying that the quantile estimates are accurate. This is true for both the method of Section 3.1 and 3.2. This conclusion is also supported by the coverage analysis of the quantiles (see Figure 4). For clarity we show the confidence interval of the coverage for selected quantiles only. The coverage of almost all quantiles is close to the expected coverage of 0.95. Only the extreme quantiles, $q \to 1$ in the exponential case and $q \to 0$ and $q \to 1$ in the normal case, are significantly smaller than expected (see Figure 4(c) and 4(e)). This shows that the method of Section 3.2 has better performance on extreme quantiles even though we are choosing the general estimator $\hat{q}_i$. This method uses a pool of observations that grows over time and therefore takes advantage of the asymptotic convergence of the estimator. In contrast to this is the method of Section 3.1. Here, the sample size is fixed and given by the number of parallel replications. This makes the use of the more specialized estimates $\hat{q}_i^{(e)}$ and $\hat{q}_i^{(g)}$ necessary.

4.2 Time Series and Queueing Models

The output processes of this section have higher complexity. However, their steady state properties are still analytically tractable. They are autocorrelated and have an initial transient phase. The CDF of the first and second example are normal and exponential, respectively. The CDF's of the third and fourth example are not covered by any of the special cases. Thus, these output processes test the full spectrum of the new estimation methods.



(a) Mean of order statistics: $F_{X_i}(x) = U(x;0,1)$

(b) Pooling: $F_{X_i}(x) = U(x;0,1)$

(c) Mean of order statistics: $F_{X_i}(x) = Exp(x;1)$

(d) Pooling: $F_{X_i}(x) == Exp(x;1)$

(e) Mean of order statistics: $F_{X_i}(x) == N(x;0,1)$

(f) Pooling: $F_{X_i}(x) == N(x;0,1)$

Figure 4: Coverage (ordinate) of the q-quantile (abscissa).

Example 1 is a geometrical ARMA process defined by

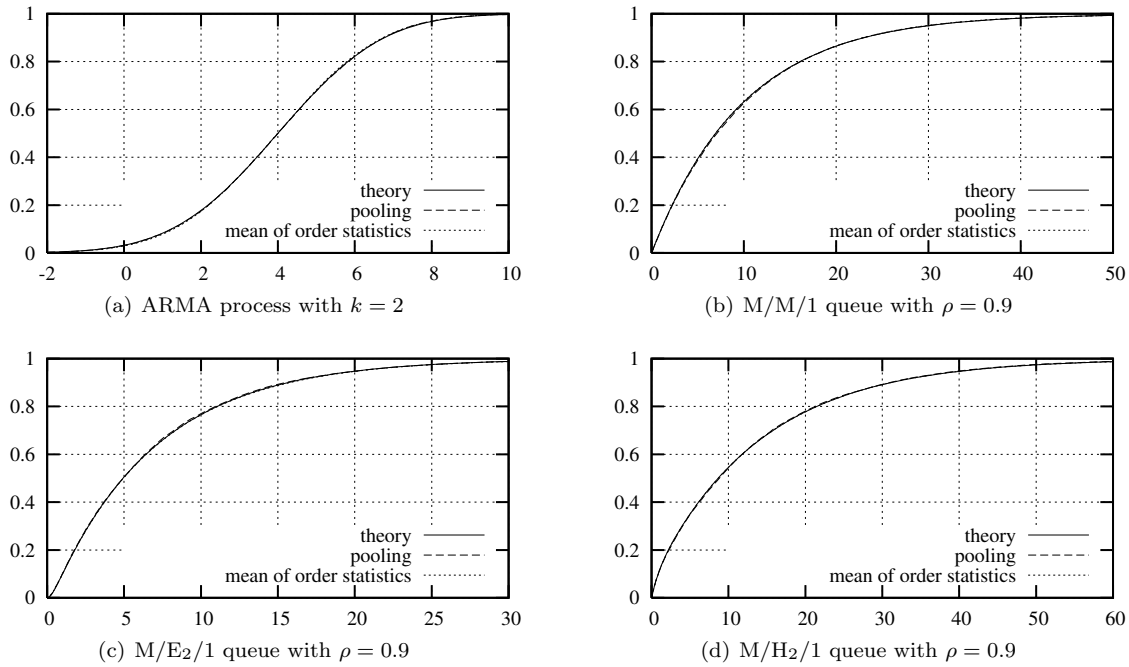
$$X_i = 1 + \epsilon_i + \sum_{j=1}^{k} \frac{1}{2^j}(X_{i-j} + \epsilon_{i-j}),$$

where ϵ_i is a Gaussian white noise process, thus, $F_{X_\infty}(x)$ has a normal distribution. $\mathrm{E}[X_\infty] = 4$ and $\mathrm{Var}[X_\infty] = 117/25$ is valid for $k = 2$ and $F_{X_\infty}(x) = \mathrm{N}(x;4,117/25)$ follows. We chose $X_i = 0$ for $i \leq 0$, which influences the length of the transient phase but not the steady state behavior.

Example 2 is the response time $\{X_i\}$ of an M/M/1 queue with service rate $\mu = 1$ and arrival rate $\lambda = 0.9$ so that $\rho = 0.9$. Here, we expect $F_{X_\infty}(x) = 1 - e^{-x\mu(1-\rho)}$. The coefficient of variation of the service time is 1.

Example 3 is the response time $\{X_i\}$ of an $\mathrm{M/E_2/1}$ queue with $\mu = 1/0.45$ and $\lambda = 1$ so that $\rho = 0.9$. The service time is given by a 2 stage Erlang distribution. $F_{X_\infty}(x)$ can be calculated by inverting the Laplace-Stieltjes transform of the response time distribution using Maple. Here, the coefficient of variation of the service time is $1/\sqrt{2}$.

Example 4 is the response time $\{X_i\}$ of an $\mathrm{M/H_2/1}$ queue, where the service time is given by a 2



(a) ARMA process with $k = 2$

(b) M/M/1 queue with $\rho = 0.9$

(c) M/E$_2$/1 queue with $\rho = 0.9$

(d) M/H$_2$/1 queue with $\rho = 0.9$

Figure 5: Known and estimated CDFs.

phase hyperexponential distribution. For a squared coefficient of variation equal to 2 with $\rho = 0.9$ we set $\lambda = 1$, $\mu_1 \approx 0.4696$, $\mu_2 \approx 1.7526$ and the probability of choosing μ_1 is ≈ 0.2113.

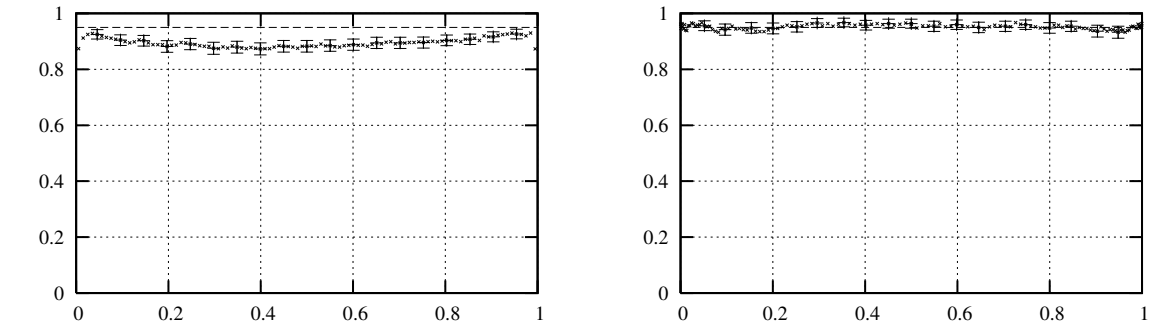
We started all simulations with an empty queue and no customer in service. In Figure 5 we can see the expected CDF compared with the estimated CDFs from the *mean of order statistics* (Section 3.1) and *pooling* (Section 3.2) (methods) for all examples. The theoretical graphs are barely distinguishable from the estimated graphs. A Q-Q plot of those graphs confirmed this assumption.

To eliminate any effects, due to simultaneous estimation of multiple quantile sequential coverage analysis was done for every estimated quantile in separate experiments. The results are presented in Figure 6. In all examples and for all quantiles the expected coverage is 0.95. We can see that the performance of the quantile estimation by pooling is very good. The coverage for all quantiles in all examples is around 0.95. This shows that estimates are approximately unbiased and the estimated confidence intervals have an appropriate size. For all experiments we used the estimate $\hat{q}_i$, as defined in Section 2.1.

The coverage of the quantiles estimated by the mean of order statistics is almost as expected for Examples 1 and 2, (Figures 6(a) and 6(c)). The coverage is significantly smaller than 0.95 for extreme quantiles. However, here we used our knowledge of the form of the distribution function and applied the more specialized estimates $\hat{q}_i^{(e)}$ and $\hat{q}_i^{(g)}$, as defined in Section 2.1. In Figures 6(e) and 6(g) the coverage of the mean of order statistics for Example 3 and 4 is much poorer. Even non extreme quantiles show a coverage significantly smaller than 0.95. For these examples none of the estimators $\hat{q}_i$, $\hat{q}_i^{(e)}$ and $\hat{q}_i^{(g)}$ are optimal. The bad coverage appears to be caused by the constant value of the sample size p. Here, all order statistics provide a slightly biased estimate. The use of specialized estimators for the calculation of the mean does not eliminate this bias.

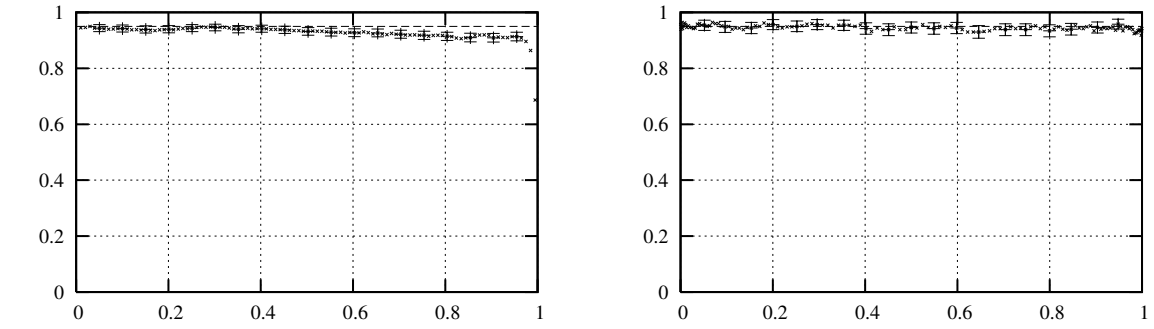
To support this we draw the empirical distribution of the mean of the 5th order statistics for Example 3 in Figure 7(a). The solid arrow marks the position of the overall mean and the dashed arrow marks the expected value. The distribution is not centered around the expected value, thus, the estimator is biased. However, the difference between the expectation and the mean $F_X^{-1}(x) - \hat{F}_X^{-1}(x) \approx 0.0082$ is small. This explains why the distributions in Figure 5(c) are almost identical and it shows that it is the constant and relatively small sample size p that causes the bias.

To show that the estimation of the variance by batching or spectral analysis is not the problem we did another series of experiments. In this case we used an independent and identically distributed output



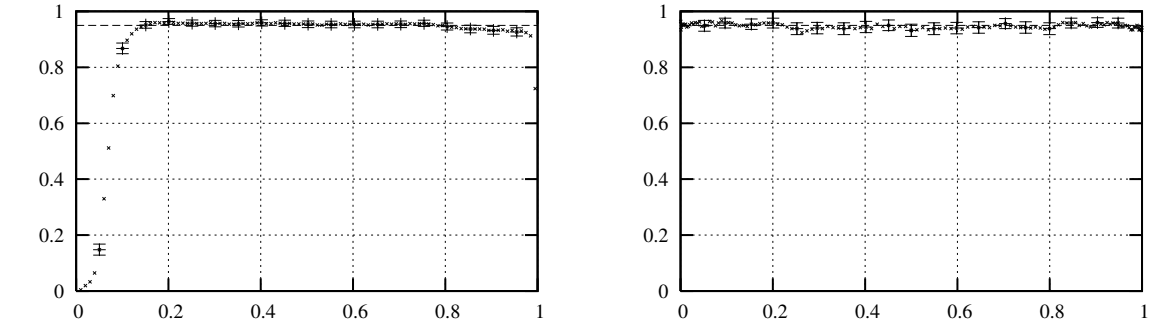
(a) Mean of order statistics: ARMA process with $k = 2$

(b) Pooling: ARMA process with $k = 2$



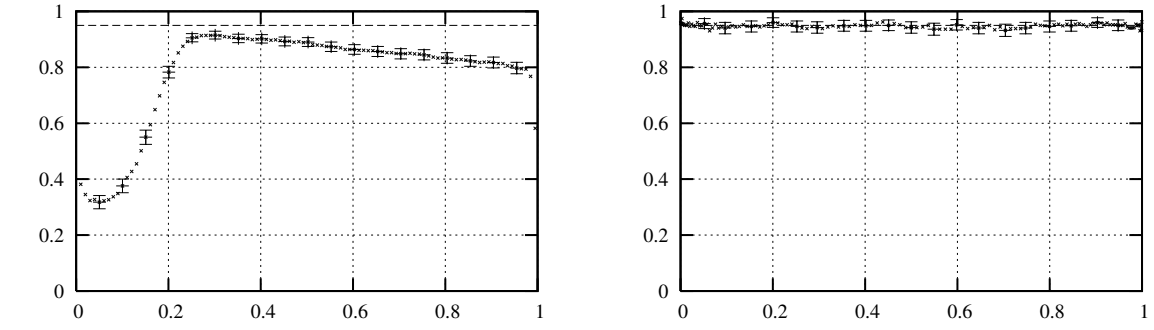
(c) Mean of order statistics: M/M/1 queue with $\rho = 0.9$

(d) Pooling: M/M/1 queue with $\rho = 0.9$



(e) Mean of order statistics: M/E_2/1 queue with $\rho = 0.9$

(f) Pooling: M/E_2/1 queue with $\rho = 0.9$

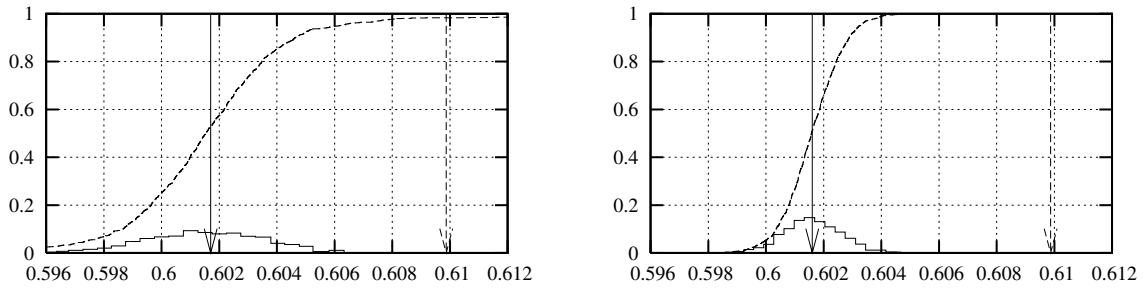


(g) Mean of order statistics: M/G/1 queue with $\rho = 0.9$

(h) Pooling: M/G/1 queue with $\rho = 0.9$

Figure 6: Coverage of the q-quantile.

process. Here, the data is drawn directly from the steady state distribution of Example 3. The empirical distribution of the mean of the 5th order statistics is depicted in Figure 7(b). The distribution of this process has smaller variance than Figure 7(a), yet the bias still remains approximately 0.0082. This shows that batching or spectral analysis does not influence the result. The constant sample size $p = 99$



(a) $M/E_2/1$ queue with $\rho = 0.9$ (b) Independent and identically distributed process

Figure 7: Empirical CDF of the mean of the 5th order statistic, where $p = 99$.

is the only source of bias.

5 CONCLUSIONS

We have considered two methods for sequential estimation of steady-state quantiles from simulation output data generated by multiple replications. The methods are particularly suited for simulation in a multiple-replications-in-parallel scenario [4], similar to that implemented in for example in Akaroa2, a controller of sequential simulation, that supports automated on-line analysis of simulation output data. Akaroa2 can support execution of arbitrary DES programs and is linked with a number of simulation packages, including the network simulation package NS2. Akaroa2 can be freely downloaded for academic purposes from *http://www.akaroa.canterbury.ac.nz*.

The first method, Means of Order Statistics, uses means of order statistics over p independent replications, as they evolve in time. Thus, the sample size remains constant. The methods provides good quantile estimates only if at least the general form of the distribution is known. In other cases the estimates are biased because the sample size p is relatively small. We tested up to $p = 200$ parallel replications, which seems to be a current practical limit. In the immediate future it appears unlikely to lead to better results than our second method, based on Pooling Spaced Observations, which uses one large pool P of (almost) independent observations. The experimental results show that the latter method is able to provide valid estimates for a wide range of quantiles. It is robust and suitable for automated on-line analysis of simulation output data, with no previous knowledge of the simulated processes. In addition it is relatively immune from the effects of processor or network communications failure, unlike the first method. The method will be implemented in a new version of Akaroa2.

Our experimental results show also that the methods can be used for estimation of steady state distributions. In the cases considered, the estimated cumulative distribution functions are almost indistinguishable from the expected ones. Small statistical errors are guaranteed by sequential quantile analysis.

References

[1] D. J. Daley, "The serial correlation coefficients of waiting times in a stationary single server queue," *The Journal of the Australian Mathematical Society*, vol. 8, no. 4, pp. 683–699, 1968.

[2] N. Blomqvist, "The covariance function of the m/g/1 queueing system," *Skandinavisk Aktuarie Tidskrift*, vol. 50, pp. 157–174, 1967.

[3] N. Blomqvist, "Estimation of the waiting-time parameters in the gi/g/1 queueing system," *Skandinavisk Aktuarie Tidskrift*, vol. 52, pp. 125–136, 1969.

[4] D. McNickle, K. Pawlikowski, and G. Ewing, "Akaroa2: A controller of discrete-event simulation which exploits the distributed computing resources of networks," *Proceedings 24th European Conference on Modelling and Simulation (ECMS 2010)*, pp. 104–109, 2010.

[5] M. Eickhoff, "Sequential Analysis of Quantiles and Probability Distributions by Replicated Simulations.," phd thesis, Department of Computer Science and Software Enginering, University of Canterbury in Christchurch, New Zealand, 2008.

[6] E. J. Chen and W. D. Kelton, "Estimating steady-state distributions via simulation-generated histograms," *Computers & Operations Research*, vol. 35, pp. 1003–1016, 2008.

[7] E. Chen, "Metamodels for estimating quantiles of systems with one controllable parameter," *Simulation*, vol. 85, no. 5, p. 307, 2009.

[8] H. A. David and H. N. Nagaraja, *Order Statistics, 3rd Edition*. John Wiley & Sons, Inc., 2003.

[9] B. C. Arnold and N. Balakrishnan, *Lecture Notes in Statistics: Relations, Bounds and Approximations for Order Statistics*. Springer, 1989.

[10] W. Conover, *Practical Nonparametric Statistics*. New York: John Wiley & Sons, Inc., 1999.

[11] D. C. Wood and B. Schmeiser, "Overlapping batch quantiles," *Proceedings of the 1995 Winter Simulation Conference*, pp. 303–307, 1995.

[12] F. N. David and N. L. Johnson, "Statistical treatment of censored data," *Biometrika*, vol. 41, pp. 228–240, June 1954.

[13] M. G. Kendall, "Note on the distribution of quantiles for large samples," *Supplement to the Journal of the Royal Statistical Scociety*, vol. 7, no. 1, pp. 83–85, 1940.

[14] G. Berry and P. Armitage, "Mid-p confidence intervals: A brief reviem," *The Statistician*, vol. 44, no. 4, pp. 417–423, 1995.

[15] P. Sen, "On the bahadur representation of sample quantiles for sequences of ϕ-mixing random variables," *Journal of Multivariate Analysis*, vol. 2, no. 1, pp. 77–95, 1972.

[16] D. L. Iglehart, "Simulating stable stochastic systems, vi: Quantile estimation," *Journal of the Association for Computer Machinery*, vol. 23, pp. 347–360, April 1976.

[17] A. F. Seila, "A batching approach to quantile estimation in regenerative simulations," *Management Science*, vol. 28, pp. 573–581, May 1982.

[18] P. Heidelberger and P. Lewis, "Quantile estimation in dependent sequences," *Operations Research*, vol. 32, pp. 185–209, February 1984.

[19] R. Jain and I. Chlamtac, "The P^2 algorithm for dynamic calculations of quantiles and histograms without storing observations," *Communications of the ACM*, vol. 28, pp. 1076–1085, October 1985.

[20] E. J. Chen and W. D. Kelton, "Simulation-based estimation of quantiles," *Proceedings of the 1999 Winter Simulation Conference*, pp. 428–434, 1999.

[21] P. Heidelberger and P. D. Welch, "A spectral method for confidence intervall generation and run length control in simulations," *Communications of the ACM*, vol. 24, pp. 233–245, April 1981.

[22] G. S. Fishman and L. S. Yarberry, "An implementation of the batch means method," *INFORMS Journal on Computing*, vol. 9, pp. 296–310, Summer 1997.

[23] A. N. Avramidis and J. R. Wilson, "Correlation-induction techniques for estimating quantiles in simulation experiments," *Proceedings of the 1995 Winter Simulation Conference*, pp. 268–277, 1995.

[24] A. N. Avramidis and J. R. Wilson, "Correlation-induction techniques for estimating quantiles in simulation experiments," *Operations Research*, vol. 46, pp. 574–591, July-August 1998.

[25] X. Jin, M. C. Fu, and X. Xiong, "Probabilistic error bounds for simulation quantile estimators," *Management Science*, vol. 14, pp. 230–246, February 2003.

[26] K. E. E. Raatikainen, "Simultaneous estimation of several percentiles," *SIMULATION*, vol. 49, pp. 159–164, October 1987.

[27] K. E. E. Raatikainen, "Sequential procedure for simultaneous estimation of several percentiles," *Transactions of the Society for Computer Simulation*, vol. 7, no. 1, pp. 21–44, 1990.

[28] K. E. E. Raatikainen, "Simulation-based estimation of proportions," *Management Science*, vol. 41, pp. 1202–1223, July 1995.

[29] S. Hashem and B. W. Schmeiser, "Algorithm 727 quantile estimation using overlapping batch statistics," *ACM Transactions on Mathematical Software*, vol. 20, pp. 100–102, March 1994.

[30] D. C. Wood and B. Schmeiser, "Consistency of overlapping batch variances," *Proceedings of the 1994 Winter Simulation Conference*, pp. 316–319, 1994.

[31] E. J. Chen and W. D. Kelton, "Quantile and histogram estimation," *Proceedings of the 2001 Winter Simulation Conference*, pp. 451–459, 2001.

[32] E. J. Chen, "Two-phase quantile estimation," *Proceedings of the 2002 Winter Simulation Conference*, pp. 447–455, 2002.

[33] E. J. Chen and W. D. Kelton, "Quantile and tolerance-interval estimation in simulation," *European Journal of Operations Research*, vol. 168, pp. 520–540, 2006.

[34] D. E. Knuth, *The Art of Computer Programming*. Addison Wesley, 1998.

[35] E. J. Chen and W. D. Kelton, "Empirical evaluation of data-based density estimation," *Proceedings of the 2006 Winter Simulation Conference*, pp. 333–341, 2006.

[36] M. Eickhoff, D. McNickle, and K. Pawlikowski, "Detecting the duration of initial transient in steady state simulation of arbitrary performance measures," in *Proceedings of the 2nd international conference on Performance evaluation methodologies and tools*, ValueTools '07, (ICST, Brussels, Belgium, Belgium), pp. 42:1–42:7, 2007.

[37] G. S. Fishman, "Grouping observations in digital simulation," *Management Science*, vol. 24, pp. 510–521, January 1978.

[38] M. Eickhoff, "Steady state quantile estimation," *Proceedings of the 13th GI/ITG Conference on Measurement, Modeling and Evaluation of Computer and Communication Systems (MMB'06)*, pp. 155–171, 2006.

[39] K. Pawlikowski, "Steady-state simulation of queueing processes: a survey of problems and solutions.," *ACM Computing Surveys*, vol. 22, pp. 123–170, June 1990.

[40] F. S. Swed and C. Eisenhart, "Tables for testing randomness of grouping in a sequence of alternatives," *The Annals of Mathematical Statistics*, vol. 14, pp. 66–87, March 1943.

[41] A. Wald and J. Wolfowitz, "Statistical tests based on permutations of the observations," *The Annals of Mathematical Statistics*, vol. 15, pp. 358–372, December 1944.

[42] J. C. Strelen, "The accuracy of a new confidence interval method," *Proceedings of the 2004 Winter Simulation Conference*, pp. 654–662, 2004.

[43] E. J. G. Pitman, "Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test," *Supplement to the Journal of the Royal Statistical Society*, vol. 4, no. 2, pp. 225–232, 1937.

[44] K. Pawlikowski, D. McNickle, and G. Ewing, "Coverage of confidence intervals in sequential steady-state simulation," *Journal of Simulation Practise and Theory*, vol. 6, no. 3, pp. 255–267, 1998.